Compres sing gr aphs with s emantic s tr ucture

Elaine Ange lino

Abstract. Graph com pres sion s chem es are bas ed on finding an orde ring or c lus te ring of no de s that plac es s imilar no des close to one anothe r. Exis ting algorithm s for We b graphs and s o c ial networks te nd to only c ons ide r no des and e dge s, whe re e ach no de com es with som e identifie r (a name or numb er), and edges are eithe r direc ted or undire cte d but othe rwise unlab e le d. Howeve r, real world graphs te nd to come with additional s em antic inform ation that is typic ally ignore d by the se com pres sion algorithms . For example , no des corre sp onding to individuals in a so cial ne twork are usually ass o ciated w ith struc tured us er profiles . Here , we b e gin to s tudy whethe r this kind of s em antic struc ture c an b e le verage d to pro duce compres sionfriendly orderings . We e mphas iz e that this additional s em antic s tructure c ome s “for free” in the s ense that it is usually alre ady c olle cted and stored by the curators of the se graph datas ets. Our preliminary re sults show that se mantic ordering heuris tics are use ful and may b e com p etitive with othe r orde rings for s o cial networks.

1. INTROD UCTIO N As large graph datase ts are collec ted and curate d in m ore domains , there is an increas ing nee d for effic ient
graph storage. Web graphs, so c ia l networks , and protein interac tion ne tworks are just som e e xam ple s of large graphs c orres p onding to re al world data, with hundreds of thousands to millions or billions of no des , and millions to billions of edges . T he se no des and edges are often ac com panie d by rich se mantic information that range s from struc tured data (e.g. additional no de or e dge attributes) to uns tructure d text and other data (e.g. Web page c onte nt).

While most graph com pres sion sche me s c ons ide r the “core” graph data – i.e . the no des and edges of a graph – the y te nd to ignore the other se mantic data. T his is unfortunate for s eve ral reasons. (1) Graph data c urators c olle ct, s tore and analyze the graph data in the context of extra s em antic data. For example , so c ial ne tworking and com mercial s ites probably want to collec t as much use r inform ation as p ossible, for bus ines s analytic s and m arke ting applic ations. Meaningful c ompre ss ion in real syste ms requires p erformance ove r core graph data plus extra s emantic data. (2) Intuition tells us that se mantic features may corre late with graph prop ertie s. For e xam ple, no de s with s im ilar se mantic information may b e more likely to have sim ilar ne ighb orho o ds . Es p e cially s ince many graph datas ets already include extra sem antic inform ation, it would b e attractive if this existing s tructure could b e leveraged for graph c ompre ssion and other related applications , e.g. c lique- finding, community detec tion, e tc.

Graph c ompre ss ion sche me s are bas ed on finding an ordering or c lus te ring of no des that place s no des with s im ilar ne ighb orho o ds clos e to one anothe r. For example, a graph that dec omp os es into n clique s plus jus t a fe w edges b etwee n clique s can b e enco ded as n lists of no de s, one for ea ch c lique, plus a list of the extra edges . This can b e a much m ore effic ient enco ding than enume rating all e dge s, but c lique-finding is we ll- know n to b e a hard problem. Likewis e, com puting an optim al orde ring of no des has b e en shown to b e NP hard [Chie riche tti et a l. 2009].

Several effi cient orde ring he uristics provide impre ss ive com pres sion p erformance . For e xample , the B V com pres sion sche me is bas ed on the fac t that simply le xic ographic ally sorting no de s in a We b graph by URL provides a natural, com pres sion- friendly orde ring [B oldi and Vigna 2004]. This works well b ec aus e of lexicographic locality : (1) page s tha t are c los e in this ordering te nd to have s im ilar ne ighb orho o ds , and (2) many links o cc ur w ithin the sam e dom ain, and will app ear near one another under this ordering. A s imple so c ial network, whe re no des corres p ond to individuals and e dge s c orre sp ond to so cial links (e.g. ‘is friends with’, ‘works with’), do e s not by its elf s ugge st a natural ordering analogous the lexicographic orde ring for We b graphs . This motivate s the shingle ordering heuris tic , which is effi cient and yields go o d com pres sion p e rformance [Chie riche tti et al. 2009].

Howeve r, as we p ointed out, mos t graph datase ts are ac tually more than jus t no de s and e dge s. In a s o cial network, each no de has a unique ide ntifier; thes e identifie rs do not nece ss arily provide a natural compres sionfriendly ordering. At the s ame time , e ach no de is typic ally ass o ciated w ith s em antic ally rich additional data, e.g. a us er profile c ontaining data for fields s uch as ‘last name’, ‘first name ’, ‘geographic lo cation’, ‘profe ss ion’,

‘em ploye r’, ‘c olle ge’, ‘inte rests ’, etc . This struc tured inform ation can b e thought of as a vec tor of fe atures ass o ciated with e ach no de . Intuition tells us that no des may tend to have ne ighb ors with similar feature vec tors , at least over some s ubse t of fields . T his ide a plays an im p ortant role in many applications , s uch as various c luste ring and com munity dete ction algorithms for so cial networks , us er pre fe rence pre diction (e.g. the motivation for the Netflix challenge), and c oarse he uristics for sharding large graph datas ets across multiple se rve rs (e.g. partitioning use r acc ounts by geographic region).

We are ge ne rally intere sted in e xploiting struc ture d s emantic information to c ompre ss graphs. In this pap er, we evaluate the usefulnes s of rudim entary s emantic heuris tic s for obtaining c ompress ion- friendly orde rings on no de s. Our basic idea is to obtain no de orderings by lexicographically or num erically s orting ass o ciated s tructured s em antic information, e.g. s ort no des by a use r- provide d tag, or by tag followe d by times tamp. For the datase t tes ted, we obse rve c ompre ss ion p e rformance unde r sem antic ordering he uristics to b e com p etitive w ith shingle ordering heuris tics with a s ingle s hingle , and p ote ntially com parable to double shingle orderings . The se pre lim inary re sults encourage furthe r evaluation with additional graph datas ets and more sophis tic ated se mantic ordering heuris tics.

2. RELATE D WORK Graph c ompre ss ion technique s have thus b ee n deve lop ed in se veral dom ains, inc luding for We b graphs [B oldi
and Vigna 2004] and s o cial networks [Chie richetti et al. 2009]. A general approach in graph com pres sion is to determine an ordering or clustering of no des that plac es s imilar no de s clos e to one another; a no de $N_i$can the n b e e ffic ie ntly repre se nte d in terms of diff erence s with re sp ect to no de N_{i-1}. De te rmining the b e st ordering is NP-hard, but “natural” and he uristic orde rings have de monstrated go o d res ults [Chierichetti et al. 2009].

Com pres sion techniques for the We b graph provide a foundation for com pres sing so c ial ne tworks. Web graphs are highly com pres sible due to their prop erty of “le xic ographic lo c ality”; we des crib ed this ide a in § 1, and it forms the basis of the Boldi- Vigna (B V) com pres sion sche me [Boldi and Vigna 2004]. So c ial networks lack a natural ordering, but a s hingling heuris tic plac es no des w ith s im ilar neighb ors close to one another. This provide s a basis for a BV-like sche me c alle d backlinks (BL), that further take s advantage of the rec ipro cal nature of s o cial networks [Chie riche tti et al. 2009].

If the goal is to s tudy graph com pres sion for application to real syste ms, then it is imp ortant to de sign com pres se d data s tructures that mee t the ac ce ss and c omputational ne eds of re al s ystem s. This is the motivation for the rec ently de velop e d layered lab e l propagation (LLP) m etho d, which pro duce s a c ompre ss ed graph data structure that corre sp onds to a hierarchical clustering of no des [Boldi e t al. 2010]. While many graph c lus tering algorithm s tend to pro duc e one partic ularly large no de, the LLP sche me is des igned to avoid this problem . The hierarchical nature of the compres sed data structure is partic ularly attrac tive for sys te ms

+We note that real storage s ystem s for m anaging graph data are s till in their infanc y com pared to re lational database manage me nt sys tems , and m any do not e xploit the graphic al s tructure of their data. In fact, the se sys tems te nd to map graph data onto a relational fram ework. For e xample , the RDF-3X engine tre ats a graph as a single re lation of (no de , e dge , no de) triples , and s tore s thes e in a com pres se d c lus te red Btree [Neumann and We ikum 2008]. The triples are sorted lexicographically and compress ed via dictionary enco ding. RDF-3X dep ends on an e xhaus tive s et of indexes on top of this s tructure to achieve fast query p e rformance.

There is a rich lite rature of graph partitioning and clustering algorithms, which c an b e applied to graph com pres sion b ec aus e they tend to group similar no de s, e.g. [Fe de r and Motwani 1995]. A thorough review is b e yond the sc op e of this pap e r, but we note re levant work in com munity de tec tion [Flake e t al. 2000; Ino et al. 2005] and mo dularity [Newm an 2003; 2006].

3. COMPRESSION SCHEME Our c ompre ss ion s che me is ve ry similar to the backlinks (BL) compress ion sche me for so cial networks
[Chie riche tti e t al. 2009]. The re are two differe nc es that are mainly due to our choic e of datase t: (1) ins tead of a so c ial ne twork, we conside r a bipartite graph (e.g. two no de typ e s corre sp ond to do c ume nts and words, and edge s conne ct do c ume nts to the words the y c ontain), and (2) our edges are undire cte d, and s o we have no notion of link rec ipro city. By using ideas from the B L com pres sion sche me – in particular, link rec ipro c ity – we c ould eas ily extend our s chem e to enco de dire cted bipartite graphs .

Consider a bipartite graph G (N1, N2, E) whe re and N2 are the no de s of the two dis tinct typ e
N1 s, |Nand N2

n - v |. Let D b e the large st diffe rence, and let k
2 = log2

1
2 1 1

1

2

1, …, Vn

2 - v |, ..., |v n -1 2, E) whe re and N2
N1 s

1

1| = |N2

1

1

2
2

-1(minui∈n (u)

1

2

1

1

1

1

| =
|N|, and the edges in E are undirec ted and re pre se nt links b etwee n no des in N1. Supp ose we have an ordering of the no des in N; we dis cuss shingling and s emantic orde ring he uristics in § 4. Also s upp os

e we have dictionary mapp e d the no de s in Nto inte gers in { 1, ..., |N_2|}. Le t v b e a no de in N_1. We e nc o de v with the follow ing piece s of inform ation:

(1) Base in formation. The de gre e of v , deg (v). This re quire s at mos t m = $\log_2(|N_2| + 1)$ bits . (2) Copying. The pre ce ding no de u that v uses as a prototyp e. We e nc o de u as the diffe rence b etween u

and v . If u = v , we copy nothing. Otherwise , we write down deg (u) bits , each repres enting whether a neighb or of u is als o a neighb or of v .

(3) Residual edges. Now we mus t enco de the remaining ne ighb ors of v not de sc rib ed by the c opying ste p. Let the se ne ighb ors b e vand let us order the m s uch that $v_1 < ... < v_n$. We write down vand e nc o de the ga ps |v(D + 1) . Writing down k requires = log(k + 1) bits . De lta enco ding the res idual edges then require s m + + (n - 1) · k bits .

4. NODE ORDERINGS As in BV or BL com pres sion, the p erformance of our com pres sion s che me dep ends on the ordering of
the no de s. While finding an optimal ordering for s o c ial networks is NP- hard, a shingle orde ring heuris tic gives go o d re sults and is easy to com pute [Chie riche tti e t al. 2009]. We des crib e two c ompress ion- friendly orde rings: (1) a straightforward extension of the s hingling he uristic for bipartite graphs, and (2) a s imple orde ring heuristic that le verages s tructure d se mantic information asso c iate d with the no des in the graph.
We us e the s ame notation as in § 3. Consider a bipartite graph G , (N N and N_2 a b e the se t of no de s n (u) = {u_i ^{1} L e t u b e a n o d e i n N { s (u_i The no de s in N_1 n d l e t i t s n e i g h b o r s i n N $_2$

1, N 1 1
2

The no des in N1

1

are the
no des of the two distinc t typ es, |N|, and the e dge s in E are undire cted and re pres ent links b e twe en no des in N. We want to find a c ompre ss ion- friendly orde ring of the no des in N. 4.1 Shingle ordering he uris tic

} . Let s b e a random p
e rmutation of the no des in N. The shingle of u is the minimum elem ent in n (u) under s , M(u) = s)}). In prac tic e, a min-wise indp ende nt fam ily or eve n pairwis e inde p e ndent has h functions c an b e us ed for s [Bro de r et al. 1998; Chierichetti e t al. 2009].
can b e ordered by the ir shingles . If two no de s in Nshare m any of the same neighb ors in N, the y will w ith high probability b e near each other in the s hingle ordering. Notic e that the shingle orde ring give s a partial ordering of the no des in N. A shingle orde ring c an b e refined by c ombining s hingle s obtained under inde p e ndent hash functions : the ordering unde r a s ec ond shingling can b e us ed to bre ak ties in the firs t, etc.

4.2 Se mant ic ordering he uris tic Now, we consider additional structure d s emantic information ab out our graph. While it would b e very
inte resting to consider sem antic information on edges, we re stric t ourse lve s here to s em antic information on no des . Le t G (N, E ; S) b e our bipartite graph as b e fore , plus se mantic information S ass o ciated with each no de in N. We c an think of an ele ment in S as a d -tuple whos e k th e leme nt corre sp onds to the k th feature in a space of d feature s.
can b e ordered acc ording to a sort w ith re sp ec t to som e pro je ction of the ir ass o ciated feature s in S . This may involve lexicographic and/or num erical s orting, dep e nding on the nature of the se mantic feature s. For e xample , s upp os e S has thre e fields corresp onding to ‘nam e’ (string), ‘tim e’ (inte ger repre se nting sec onds s inc e the ep o ch), and ‘tag’ (s tring). Lexicographically s orting the ‘name’ field provides a 1-dimensional ordering of the no des in N. Le xic ographic ally sorting the ‘tag’ field followed by numerically sorting the ‘time’ fie ld provides a 2-dim ensional ordering of the no des in N. T he se orde rings bas ed on multi-dim ensional pro jec tions of our fe ature space are s im ilar in spirit to the m etho d of c ombining shingles desc rib ed ab ove. We note that the se parate dime ns ions in our feature ve ctor are in ge ne ral not inde p e ndent.

, N
3
Equ ivalently, we c ould en co de G(
N
, E) plus th e se mantic in for mati on S as a (d + 2) -part ite grap h. This m ay b e a useful rep resentat ion wo rth furt her c onsid erati on.

5. EXPERIME NTS Our prim ary goal is to c ompare se mantic orde ring heuris tic s to shingle ordering heuristics . We consider
se mantic orde rings bas ed on individual as well as multiple se mantic fields , and shingle orderings base d on individual and multiple shingle s. In this se ction, we de sc rib e the datase t we use for our e xp erime nts , bas eline no de orderings for comparison, and s ome im ple mentation de tails.

5.1 Data We fo cus on one partic ular datas et, a se t of twee ts from June 2009 obtained by c rawling the p opular
mic roblogging We b s ite, twitter.com [Yang and Les kovec 2010].[2]Each twe et is a s hort me ss age c ons is ting of up to 140 characte rs , and is ass o ciated with two additional piece s of data: a us ernam e and a time stamp (with se cond- level granularity). Each twee t its elf m ay contain two partic ularly interes ting kinds of us er-provided information: hashtags and the use of the '@' s ymb ol to de note a reply. A hashtag is a token (e.g. a word) that s tarts with the '#' symb ol; use rs can group and se arch for twee ts by hashtag, and so are motivated to provide se mantic ally me aningful hashtags. T he '@' symb ol followed by a us ernam e indicates that the twee t is a reply to the us er as so ciate d with use rname .

We think of this data as a bipartite graph G (T , V , E ; S) where T is the se t of no des c orres p onding to tweets and V is the s et of no des c orres p onding to tokens in the twe et vo c abulary. An e dge b etwee n t ∈ T and v ∈ V indicates that v is a toke n in twe et t ; the se edges are undirec ted. This bipartite graph repres ents a "bag of words" interpretation of the corpus of twee ts. We define the additional s em antic fe atures S on the tweets to include the as so ciated us ernam e and times tamp, as well as any toke ns that can b e interprete d as hashtags or re plies .

We summarize the bas ic prop e rties of our datase t in Table I (a). For s im plicity, we eliminate all twe ets with Unico de charac ters. This is a c rude way to get rid of twe ets , such as those written in Asian characte rs , that require differe nt lexing s chem es for toke niz ation. Namely, we only conside r twee ts that c an b e le xed by splitting the twe et s tring on white space ; our m etho d happ ens to e lim inate more twee ts than is strictly nece ss ary. We also only consider twe ets containing at leas t one token that lo oks like a hashtag (we do not consider the '# ' symb ol alone to b e a has htag). This yields ab out 1.7 million tweets (ab out 10% of the original datas et) and 2.0 million toke ns connecte d by 24 million e dge s.
We summarize the s em antic features of our datas et in Table I(b), i.e. the numb e r of unique use rs , times tamps, e xtrac te d replie s and has htags. In the c ase of multiple re plies , we e xtrac t only the first. In the cas e of multiple hashtags, we extract only the firs t two, and consider e ach as a s eparate feature .

(a) Ba sic gra ph p rop erti es.

Nota tion	Featu re	Numb er
T	tweets	1,7 31,4 02
V	tokens	2,0 73,0 37
E	lin ks	24 ,042, 218

(b) Se manti c featu res of tweets .

Featur e	Numb er
use r	42 3,616
tim esta mp	83 5,569
rep ly	1,7 31,0 72
prim ary ha shtag	13 8,854
sec onda ry hashta g	53 ,384

Tabl e I. S umm ary of (a) basi c gra ph p rop ertie s and (b) sem antic feat ures of our T witt er da taset .

5.2 Base line orderings For c omparis on purp ose s, we c ons ide r two bas eline
orde rings of the no de s in T :

(1) Ran dom orde r. This is a random p ermutation of the no de s. (2) Nat ural orde r. This is the orde r of the no de s in the original datas et, pres um ably c orre sp onding to the
orde r in w hich the twitter.com page s were crawle d. We note that orde rings base d on individual pie ces of se mantic information are often considered as a
baseline. For example, z ip co des of individuals in a s o cial network c an b e us ed to obtain a ge ographic orde ring [Chie riche tti et a l. 2009].

[2]We no te t hat this dat a is app arently n o lo nge r ava ilab le th roug h th e S NAP (S tanford Network Ana lys is Platform) web site, "a s p er reque st from Twi tter."

5.3 Impleme ntation details (1) Comp ression scheme. Given a no de ordering, to e nc o de a particular no de u , we only consider the

pre ce ding no de in the ordering for the prototyp e of u , i.e. we only use a window siz e of 1. (2) Hash fun ctio ns. We us e Num Py's numpy.random.permutation() function to generate p ermutations of

{1 , ..., |V |} for our s hingle orde ring heuris tic .

6. RESULTS In this s ec tion, we summarize our com pres sion p e rformance results, com paring shingle order heuris tic s to

se mantic order he uristics , for single orde rs , double orders, and additional orders (Table I I and Figure 1). We also include res ults for our bas eline orde rings. All c ompre ss ion res ults are give n in te rm s of bits p er link.

(a) Si ngle orde rs. Ord ering Bits / lin k

shi ngle	17 .455*
ha sht ag	17 .629
use r	18 .700
rep ly	19 .142
tim e	20 .334
na tural	20 .331
ran dom	21 .042

(b) Se manti c dou ble orders . ha sht ag tim e rep ly use r

ha sht ag - 16 .513	16 .637	16 .874	tim e 20 .270 - 20	306 20 .327
rep ly 17 .011	18 .220 - 17 .707	use r 18 .141	18 .026	18 .249 -

(c) 3- and 4-ord ers. Ord ering Bits / lin k

3-s hing le 15 .135 ha sht ag, t ime, reply 16 .501 ha sht ag, t ime, user 16 .499	
4-s hing le 14 .112 ha sht ag, t ime, reply, use r 16 .498 ha sht ag, t ime, user, reply 16 .499	

Tabl e I I. Su mma ry of com pressio n p e rfo rman ce fo r d ifferent orde rings an d combin atio ns of ord ering s. Al l re sults are re p o rted in bits p er li nk. No te t hat in t hese tab les, "ha shtag" refers to t he p rima ry hashta g, i.e. th e fir st ha shtag app ea ring in a tweet. (a) Re sults for singl e ord ering s. * This result fo r the singl e shin gle orderi ng re presents the b es t from 1 0 ind ep end ent tria ls (m ean = 1 7.690 bits/ link, st anda rd devia tion = 0.186 bits/ link, m ax = 18. 172 bi ts/lin k) . (b) Res ults for do uble s e mantic orderi ngs. Th e rows and colum ns co rresp ond to the fi rst an d secon d sem antic featu res use d in th e order ing. Note th at the doub le shin gle ord ering b eat s al l of these by at lea st 1 bi t/lin k, e.g . achie vin g 1 5.13 5 b its/l ink w ith two par ticul ar ra ndom ly cho sen shing les. Also no te tha t the or derin g achieved by co mbinin g prim ary and se cond ary hashta g data g ives 17 .075 b its/l ink. (c) Orde rings tha t co mbine three or fou r shin gles o r sem antic fea tures .

6.1 Base line orderings For com parison purp ose s, we re p ort re sults for our two bas eline s in Table I I(a). Com pres sion w ith the

natural and random orders require ab out 20.3 and 21.0 bits/link, res p ec tively.

6.2 Shingle orderings

We ge ne rate 10 diff erent shinglings . Firs t, we consider e ach shingle se parately. Over the 10 trials, com pres sion p e rformance ranged from ab out 17.5 - 18.2 bits /link, with a m ean and s tandard deviation of 17.7 and 0.2 bits /link, re sp ective ly. Note that the b es t p e rformance ove r the 10 trials is re p orted in Table I I(a).

We also consider combinations of inc re asing numb ers of shingles , from 1 to 10, for a particular randomly chos en order of the s hingles. We note that we ca n obtain s ignific ant p erformance ga ins by c ombining multiple shingle s, but that the se gains dec ay quickly as additional shingle s are inc luded (Figure 1). I n our partic ular exam ple, the double s hingle orde ring give s an improvem ent of ab out 2.5 bits /link. Com pres sion w ith the 4-s hingle ordering require s only 13.847 bits/link, the 5-s hingle orde ring dec rease s this by 0.096 bits /link, and the 10- shingle ordering de creas es by an additional 0.094 bits/link. We also re p ort thes e re sults for 3- and 4-s hingle orde rings in Table I I(c).

6.3 Se mant ic orderings Firs t, we c ons ider orde rs obtaine d by sorting along the following single sem antic dime ns ions : primary hashtag,

user, time stamp, re ply (Table I I(a)). The ordering by time stamp give s c ompre ssion c omparable to the natural orde r; it turns out the twe ets we re alre ady nearly ordered by time stam p. T he othe r three s ingle sem antic orde rings are all b etter than the natural ordering. In partic ular, the has htag orde ring is the b es t, and is in fact c omparable to the single s hingle orde rings.

Next, we c ons ider all 12 orders obtained by sorting acc ording to ordered pairs chose n from the four se

mantic dim ensions ab ove (Table I I(b)). We als o te ste d com pres sion obtaine d by None of the double se mantic orderings are as go o d as the double shingle orde ring (ab out 15 bits /link), with the b e st – orde ring 5

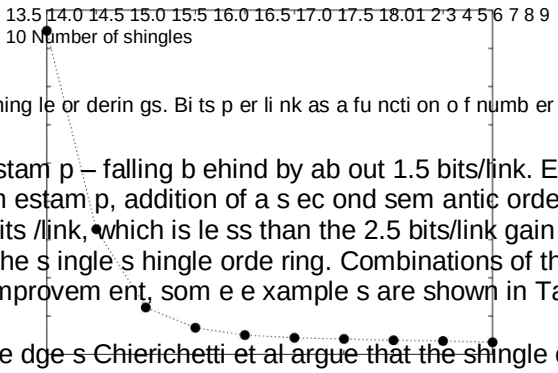
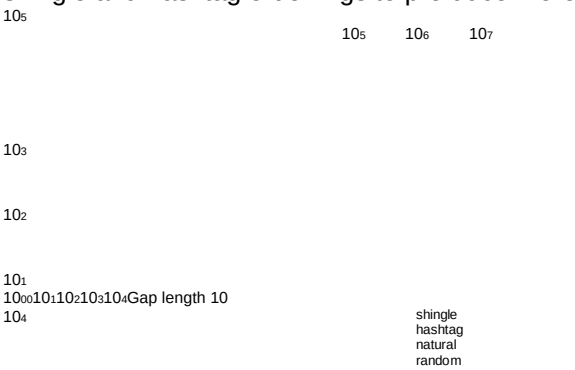


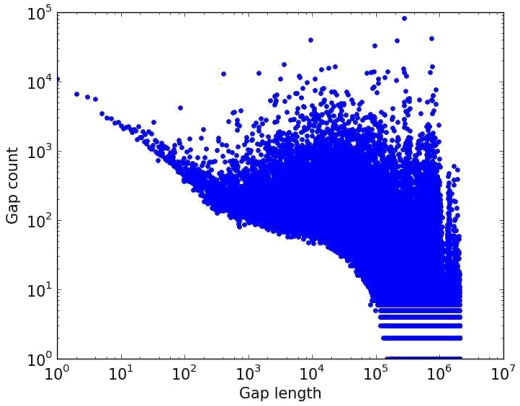
Fig. 1. Multi- shing le or derin gs. Bi ts p er li nk as a fu ncti on o f numb er of shing les.

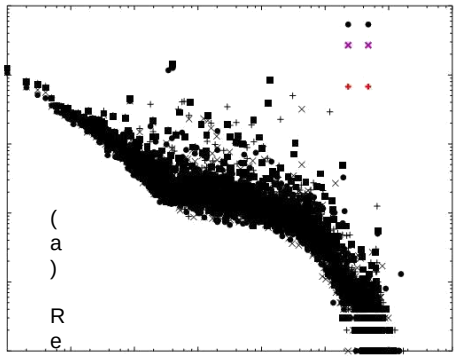
by has htag followe d by tim estam p – falling b ehind by ab out 1.5 bits/link. Exc ept for double sem antic orde rings that start s ort firs t by tim estam p, addition of a s ec ond sem antic ordering to a firs t improves p e rformance by ab out 0.5- 2.0 bits /link, which is le ss than the 2.5 bits/link gain we obtaine d by us ing the double shingle c ompare d to the s ingle s hingle orde ring. Combinations of three to four s em antic orderings only give sm all incre mental improvem ent, som e e xample s are shown in Table I I(c).

6.4 Gap lengths and re sidual e dge s Chierichetti et al argue that the shingle orde ring enables go o d compres sion by creating m ore small gaps ,
com pared to the othe r orderings . They exp e rim entally s how this by plotting the numb er of gaps ve rsus gap length for diff erent orde rings, w hich exhibit very differe nt dis tributions [Chie riche tti e t al. 2009]. We pro duce sim ilar plots in Figure 2, but obse rve only very subtle diffe re nc es b e twe en the distributions for the diffe rent orde rings. Figure 2(a) plots a repre se ntative dis tribution, the numb e r of gaps (y -axis) versus gap le ngth (x-axis) for the hashtag ordering. Following Chierichetti et al, Figure 2(b) plots a s ub-s ample d ve rs ion of the se dis tributions , for shingle (blue circles), hashtag (mage nta x's), natural (c yan s quares), and random (red c rosse s) orde rings. In particular, we we re suprise d that there was not an imme diate differe nc e b e twe en the dis tributions for the shingle (or hashtag) versus random (or natural) orderings . We exp e cted shingle and hashtag orde rings to pro duce more sm all gaps than the random and natural orde rings.



at ive distri
ring).





(a) Repr esent

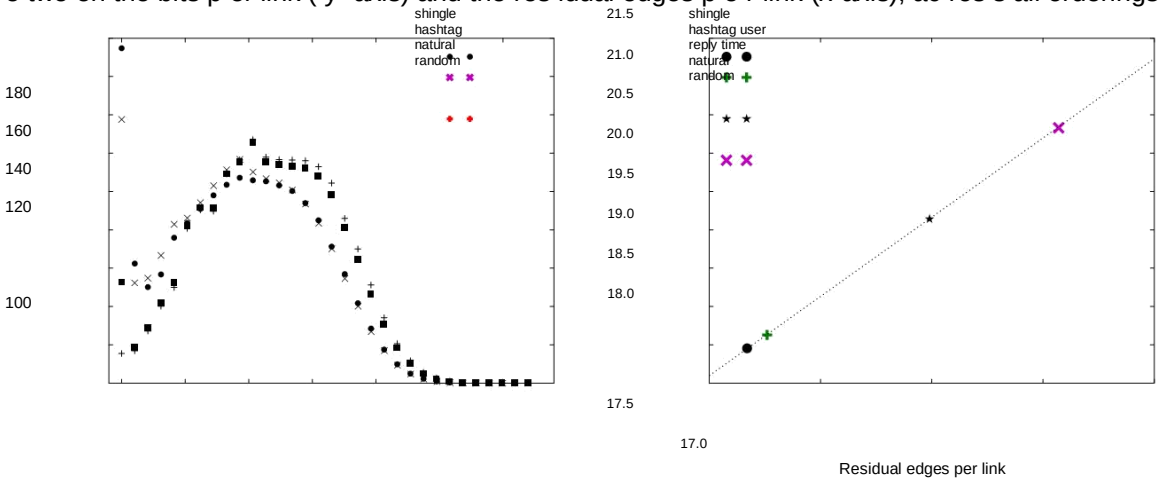
(b) Sub-sampled gap distributions.

Bits per link

Gap count

Fig. 2 . Numb er of gap s (y -axis) o f a cer tain le ngth (x -axis) , for (a) a rep resenta tive ord ering (hashta g), a nd (b) a su b-sa mpli ng of p oints for shin gle (blue circle s), h ashtag (mag enta x’s) , nat ural (cya n square s), a nd ra ndom (re d cro sses) o rderi ngs.

3Since the distribution of gap lengths did not e xplain the difference s b e twe en the orderings , we de cided to lo ok at the numb er of re sidual edges in the e nc o ding, for the diffe re nt orde rings. Figure 3(a) plots the numb e r of twe ets (in thous ands, y -axis) with a c ertain numb e r of re sidual edges (x-axis), for shingle (blue circle s), hashtag (mage nta x's), natural (c yan s quares), and random (re d c ros se s) orde rings. He re , we c an s ee that the shingle and hashtag orderings allow m ore twee ts to b e enco de d with a sm alle r numb er of re sidual edges , i.e . are b etter able to e xploit copying. Note in particular that shingling do es the b es t at ide ntifying tweets that are completely sp ecified by a copy of another twe et (whe n the numb er of res idual e dge s is ze ro).
Let the residual edges per link b e the numb er of res idual edges in the enco ding divided by the numb e r of edges in the graph. If there is no copying in the e nc o ding, then this numb er is 1; if the enco ding is achie ved by copying alone , the n it is 0. Figure 3(b) s how s that for our datase t, the re is a line ar relations hip b e twe en the bits p er link (y -axis) and the res idual edges p e r link (x-axis), ac ros s all orderings .



(b) Bit s p er lin k vs . resid ual edge s p er lin k.

Fig. 3. (a) Numb er of twe ets (i n th ousan ds, y -axis) wit h a c ertai n numb er o f re sidu al ed ges (x -axis) , for s hingl e (b lue c ircles), ha sht ag (m agenta x's) , natu ral (cyan square s), an d rando m (red crosse s) orderi ngs. (b) Bits p er lin k versus th e nu mb e r o f resi dual edge s p er li nk, for diffe rent ord ering s. Th e do tted line corre sp o nds to th e equa tion, y = 20 . 67 2x + 2. 62 8.

We sp eculate that the diff erent character of our plots in Figure (2) c ompare d to thos e obse rve d by Chierichetti et al has to do with diffe re nc es in the data that we chose to c ompre ss, name ly, that our Twitter graph repre se nts natural language data (re call that it is a bipartite graph of tweets and words in tweets) rathe r than a s o cial network or Web graph. In partic ular, re call that the numb e r of bits us ed to e nc o de each res idual e dge for a particular no de is se t by the s iz e of the large st gap. It s ee ms that this numb er is not partic ularly affe cte d by the orde ring, which do es not see m unreas onable for se ts of re sidual edges that corres p ond to se ts of c o-o cc uring words .

7. FUTURE DIRECTIONS Our study has b e en fairly limited in sc op e. For our se mantic orde ring heuris tic, we only considered sem antic
feature s of no des ; it would b e interes ting to als o conside r s emantics on edges . Als o, we only c ons idered the mos t s im ple m etho ds for us ing and c ombining se mantic fe atures for pro duc ing orderings base d on lexicographic (or num eric al) sorts. It may b e worth exploring more s ophisticate d m etho ds for combining multiple se mantic features to pro duc e an ordering or c lus te ring of no des .

Thes e real world graphs typically live in databas es , where c ompre ss ion must (1) ac company effic ient partitioning ac ross files on disk, and p os sibly s harding across multiple s ervers for ve ry large graphs , and (2) b e balanc ed w ith effi cient acc es s me tho ds for que ry exec ution, as the re are us ually trade -off s b e twe en the two. It is worth noting that the lexicographic and/or nume ric al s orts employed by our s imple sem antic orde ring heuris tic s may corres p ond to a natural layout of data within a database and/or us eful database indexe s (which can b e e fficie ntly compres sed b e cause they are s orte d). As m entioned earlier,

it is imp ortant

3"Retweetin g" – wh en a user c opie s ano ther user's twe et – is ap parently p op ular amo ng T witt er use rs.
7

to evaluate graph c ompre ss ion scheme s not only with re sp ect to s torage p e rformanc e, but also in the c ontext of real s ys tems with data acc es s nee ds .

We have only studied one particular datas et, a c orpus of s hort m es sage s that we tre ated as “bags of words ” and represe nted as a bipartite graph. While our datas et is not standard in the lite rature for e valuating graph com pres sion p erformance , it allowe d us to de monstrate e nc ouraging res ults for our se mantic orde ring heuristic . As additional side e ffec ts, this datas et also gave us an e xcuse to conside r com pres sion for bipartite graphs and in particular, graphs ass o c iated with natural language .

8. CONCLUSIONS The We b graph’s prop erty of le xic ographic lo cality allows an e ffic ie nt orde ring heuris tic – lexicographically
sorting the no de s by URL, i.e. no de “name ” – that can b e exploited by a s chem e like B V c ompre ss ion. We e ss entially extende d the notion of lexicographic lo c ality to include any s emantic fe atures on the graph, such as additional struc tured attribute s ass o ciated with no de s. T his idea is e sp ec ially us eful for graphs whose no de names alone do not provide a natural ordering, if the no de s are asso c iate d w ith othe r sem antic information that c orres p onds to s ome notion of lo c ality, s uch as so cial networks , w hos e no de s c orre sp ond to individuals who each have a structure d use r profile .

9. ACKNOWLEDGEMEN TS The author would like to thank M. Seltze r for sugges ting e dge lab e ls as a he uristic for c lutering no des in a
graph, V. Kanade for providing re fe rece s ab out c ommunity de tec tion and influe nc ing the language in § 4.2, and M. Mitzenmacher for many fun and rele vant readings, le cture s, as signm ents, comm ents, e tc.

REFERENCES Bol di, P. , Rosa, M., San tini , M. , an d Vigna , S. 20 10. Layer ed lab el pr opag ation : A mu ltires oluti on co ord inate -fr ee orde ring

for co mpre ssing so cial networks. CoRR abs/1011.5425. Bol di, P. a nd Vi gna, S. 20 04. T he we bgrap h fram ework i: c omp ressio n techn iques. In WWW ’04: Proceedin gs of the 13th

in ternation al conf erence on World Wide Web. ACM, New Yor k, NY, US A, 595 –602 . Brod er, A. Z . , Cha rikar, M. , Fr ieze, A. M. , an d Mi tzenmacher, M. 19 98. Min-w ise In dep end ent Pe rmutat ions. Jou rn al

of Computer and System S cien ces 60 , 3 27–3 36. Chi er ichetti , F . , Kumar , R. , Lattanzi , S. , Mitzen mach er, M. , Pan con esi , A. , an d Ragh avan, P. 20 09. On com pressin g

so cial networks. In KDD ’ 09: P roceedin gs of the 15th ACM S IGKDD intern ational conferen ce on Know ledge discovery an d data min in g. ACM, New Yor k, NY, US A, 219 –228 .

Fed er , T. and Mot wa ni, R. 19 95. Cl ique p artiti ons, grap h co mpre ssion and sp eed ing-u p a lgorit hms. J. Comput. S yst. Sci. 51, 26 1–27 2.

Fl ake, G. W. , Lawr en ce, S. , an d G iles, C. L. 20 00. Effi cient i dentifi catio n of web co mmuni ties. In Proceedin gs of the sixth ACM S IGKDD in ternati on al con feren ce on Know ledge discovery and data mi nin g. K DD ’0 0. ACM, New Yo rk, NY, USA, 15 0–16 0.

Ino, H. , Kud o, M. , an d Na ka mu ra, A. 20 05. Par tition ing of web g raphs by co mmun ity top olo gy. In Proceedin gs of the 14th in ternation al conf erence on World Wide Web. WWW ’ 05. ACM, New Yo rk, NY, USA, 66 1–66 9.

Neumann , T. and Weikum, G . 20 08. Rdf-3x: a risc-s tyle engin e for rd f. Proc. VLDB En dow. 1, 1, 647– 659. Newman , M. E. J. 20 03. Fast a lgori thm fo r det ectin g co mmuni ty st ructu re in networks. Newman , M. E. J. 20 06. Mo d ulari ty an d co mmuni ty str uctu re in n etworks. Proceedin gs of the Nati on al Academy of S ci -

ences 103, 23 (Ju ne), 8577 –85 82. Ya ng, J. and Leskov ec, J. 20 10. 47 6 m illion twitte r twe ets. ht tp:/ /sna p.st anfo rd.e du/d ata/ twit ter 7.ht ml .